Clustering Similar Neighborhood in New York Cities

Ashfaqul Alam Joarder

May 2020

1. Introduction

1.1 Background

The York City's demographics show that it is a large and ethnically diverse metropolis. It is the largest city in the United States with a long history of international immigration. New York City was home to nearly 8.5 million people in 2014, accounting for over 40% of the population of New York State and a slightly lower percentage of the New York metropolitan area, home to approximately 23.6 million. Over the last decade the city has been growing faster than the region. The New York region continues to be by far the leading metropolitan gateway for legal immigrants admitted into the United States.

1.2 Problem

Finding identical neighborhoods in different cities in order to help provide a perception of similar neighborhoods which may provide with a great deal of insights in order to make a decision of choosing a neighborhood that is far away, yet somewhat feels like home.

Throughout its history, New York City has been a major point of entry for immigrants; the term "melting pot" was coined to describe densely populated immigrant neighborhoods on the Lower East Side. As many as 800 languages are spoken in New York, making it the most linguistically diverse city in the world. English remains the most widely spoken language, although there are areas in the outer boroughs in which up to 25% of people speak English as an alternate language, and/or have limited or no English language fluency. English is least spoken in neighborhoods such as Flushing, Sunset Park, and Corona. With it's diverse culture, comes diverse food items. There are many restaurants in New York City, each belonging to different categories like Chinese, Indian , French etc.

2. Data Acquisition and Cleaning

2.1 Data Sources

This project works with two sets of data. The first dataset consists of New York's different neighborhoods and their respective geometric coordinates, which can be found here.

2.2 Generating the Data

Describe the data that you will be using to solve the problem or execute your idea. Remember that you will need to use the Foursquare location data to solve the problem or execute your idea. You can absolutely use other datasets in combination with the Foursquare location data. So make sure that you provide adequate explanation and discussion, with examples, of the data that you will be using, even if it is only Foursquare location data.

New York City data that contains list Boroughs, Neighborhoods along with their latitude and longitude. Data source : https://cocl.us/new_york_dataset Description : This data set contains the required information. And we will use this data set to explore various neighborhoods of new york city. Indian

restaurants in each neighborhood of New York city. Data source: Foursquare API Description : By using this api we will get all the venues in each neighborhood.

We can filter these venues to get only Indian restaurants. Geospacer Data source : https://data.cityofnewyork.us/City-Government/Borough-Boundaries/tqmj-j8zm Description : By using this geo space data we will get the New York Borough boundaries that will help us visualize choropleth map.

2.2 Data Cleaning

The first data source in the described link is in json format. It initially consisted of many different classes of data. Upon examining them, the data that we are interested in was found under 'features' category. Further formatting of the json data finally resulted in a dataframe that consists of 4 columns, namely: Borough, Neighborhood, Latitude and Longitude.

But the problem with this dataframe was, it has some values under the column 'Borough' which were not assigned in the first place. So, the rows with no assigned value in the 'Borough' column were dropped. Another problem was there were a few rows in the 'Neighborhood' column that too had no values assigned to it. As a solution, the value

2.3 Approach

Input:

- Collect the New York city data from https://cocl.us/new_york_dataset
- Using Foursquare API, we will find all venues for each neighborhood.
- Filter out all venues that are Indian Restaurants.
- Find rating, tips and like count for each Indian Restaurants using Foursquare API.
- Using rating for each restaurant, we will sort that data.
- Visualize the Ranking of neighborhoods using folium library(python)

Queries that can be answered using above dataset

- What is best location in New York City for Diversified Cuisine?
- Which areas have potential Asian and Arabian Restaurant Market?
- Which all areas lack Diversified Restaurants?
- Which is the best place to stay if customers prefer specific Cuisine?

2.4 Analysis: Required Libraries

- Pandas and numpy for handling data.
- Request module for using FourSquare API.
- Geopy to get co-ordinates of City of New York.
- Folium to visualize the results on a map

2.5 Feature Selection

Now that we have obtained the different neighborhoods and their respective geometric coordinates for the city of New York and Toronto, it is time to come up with different venues that the different venues have to offer.

Foursquare API provides with an access to an enormous database consisting of venues from all around the world including rich variety of information such as addresses, tips, photos and comments. Having signed up for a Foursquare developer, using the Client ID and Client Secret, it is possible to make API requests in order in order to retrieve venue information.

By feeding a function with Neighborhood name and its geometric coordinates, using Foursquare API different venues (Restaurants, Coffee shops, etc) were extracted. After performing One-Hot-Encoding and grouping together the rows by neighborhoods, the NY dataset and Toronto dataset seemed to share 250 features. Both the dataframe were combined into a single dataframe in order to perform clustering operation.

2.4 Dimensionality Reduction

Principal Component Analysis is a dimension-reduction tool that can be used to reduce a large set of variables to a small set that still contains most of the information from the large set. Principal component analysis (PCA) is a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components. Before diving into clustering operation, first we performed dimensionality reduction using Principal Component Analysis on the dataframe in order to reduce the number of dimensions.

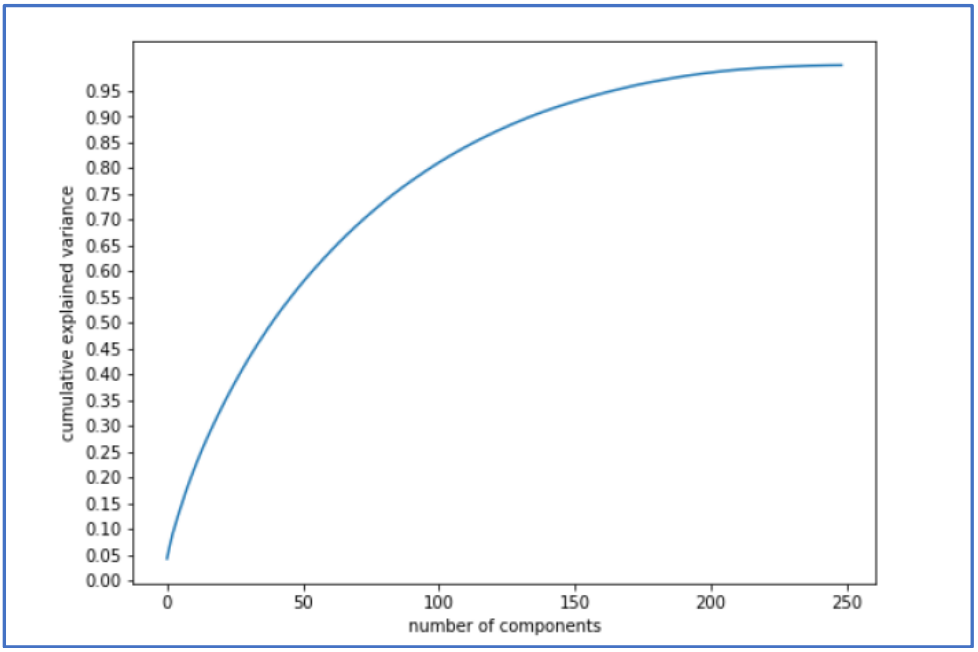

Fig : Selection of Number of Principal Component

Having performed PCA, the number of features was reduced to 150 from 250 yet retaining the maximum variance of the dataset.

3 Methodology

The goal of this project is to group together the similar neighborhoods in the city of New York and Toronto. Since the dataset is unlabeled i.e. unsupervised, this

3.1 Determining Optimal Cluster Number

K-means is a simple unsupervised machine learning algorithm that groups a dataset into a user-specified number (k) of clusters. The algorithm is somewhat naive--it clusters the data into k clusters, even if k is not the right number of clusters to use. Therefore, when using k-means clustering, users need some way to determine whether they are using the right number of clusters. One method to validate the number of clusters is the elbow method. The idea of the elbow method is to run k-means clustering on the dataset for a range of values of k (say, k from 1 to 20), and for each value of k calculate the sum of squared errors (SSE). Then, plot a line chart of the SSE for each value of k. If the line chart looks like an arm, then the "elbow" on the arm is the value of k to be used. The idea is that we want a small SSE, but the SSE tends to decrease toward 0 as we increase k (the SSE is 0 when k is equal to the number of data points in the dataset, because then each data point is its own cluster, and there is no error between it and the center of its cluster). So our goal is to choose a small value of k that still has a low SSE, and the elbow usually represents where we start to have diminishing returns by increasing k.

The second method to find our optimal cluster number is silhouette analysis. Silhouette analysis is a way to measure how close each point in a cluster is to the points in its neighboring clusters. It is a neat way to find out the optimum value for k during k-means clustering. Silhouette values lies in the range of [-1, 1]. A value of +1 indicates that the sample is far away from its neighboring cluster and very close to the cluster it is assigned. Similarly, value of -1 indicates that the point is close to its neighboring cluster than to the cluster it is assigned. And, a value of 0 means it’s at the boundary of the distance between the two clusters. Value of +1 is ideal and -1 is least preferred. Hence, higher the value better is the cluster configuration.

3.2 Random Initialization

Since, K-means incorporates a heuristic approach, it does not ensure converging at global optima at each iteration. Depending upon how the initial position of the clusters were set, it maybe converges into different local optima. In order to overcome this issue, numerous iterations on different random initializations were performed in order find the best set of convergence.

4 Results

4.1 Optimal Number of Clusters

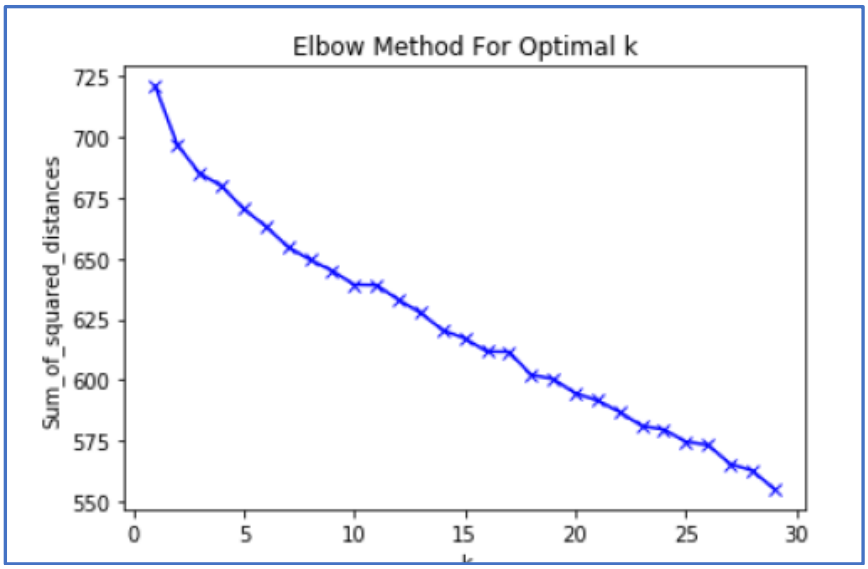

Figure 4.1 Elbow method to determine the number of K

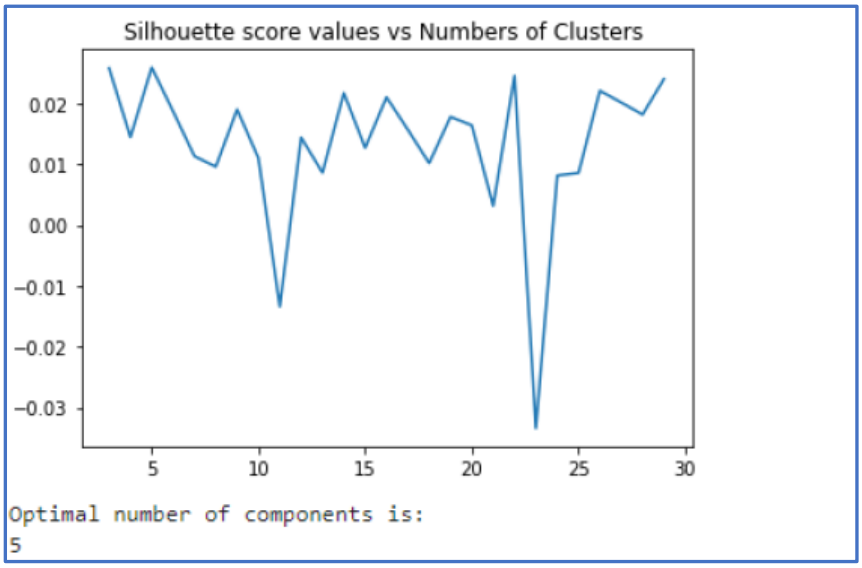

Figure 4.2 Silhouette Score to determine the number of K

The number of clusters being 5 experiences a decrease before it and a gradual regular decrease after it in figure 4.1. The Silhouette score confirms number of clusters being 5 has its peak in figure 4.2. The Silhouette Coefficient is calculated using the mean intra-cluster distance and the mean nearest-cluster distance for each sample. Figure 7 shows the elbow method, in which case, we can't see a distinctive elbow, but we can still come to a conclusion that 15 clusters would be a reasonable choice.

4.2 Visualizing the Clusters on the Map

We created folium maps to help obtain a visual perception of how the very different clusters look on the map when plotted on the map of the city of New York and Toronto.

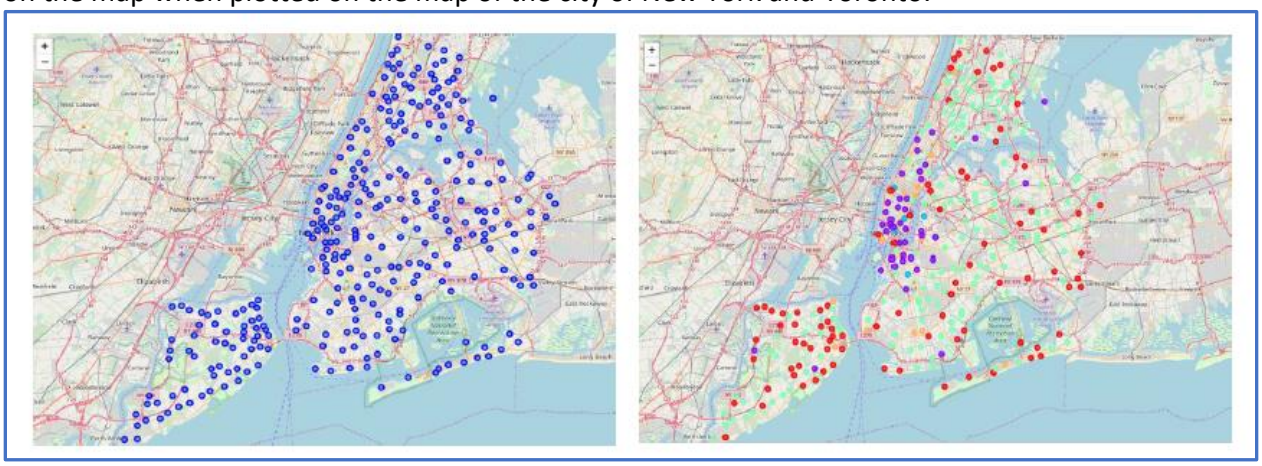

Figure 4.2 (a) Figure 4.2(b)

Figure 4.2 (a) Venues pinned on NY before Clustering (b) Venues pinned on NY after Clustering

The very different venues in the city of NY have been clustered into 5 distinctive clusters represented by different colors in the Figure 4.2 (b). Now, to comply with the initial goal of finding similar neighborhood in a different city is being visualized in the Figure 4.3 (b). The points of similar color in the Figure 4.2 (b) and again, in the Figure 4.3 (c) represent similar neighborhood in terms of venue information that we obtained from Foursquare API.

5 Discussion

Since this is an unsupervised clustering work, many different approaches can be adopted in order to achieve better results. The project was only done on the zip codes of New York, each having 150 features, even after performing dimensionality reduction. Having more samples may result in a better clustering. For instance, for the outliers that are being observed on the maps could be defined by using DBSCAN algorithm.

Having dealt with location data on a deeper level, for instance at neighborhood level may result in better grouping of similar data points which eventually may result is better clustering. The study here is being ended by visualizing the data and clustering information on the map of the City of New York.

6 Conclusion

People are frequently moving into new cities. And in this ever growing world filled with technology, having a neighborhood recommendation based on location data is something to be considered basic now-a-days. And the application of neighborhood segmentation lies beyond this application too. This can serve to be an impressive tool to better organize a city resource. Furthermore, it can be used as a tool for security measurement if combined with crime data.